U. S. NAVAL SCHOOL OF AVIATION MEDICINE
NAVAL AIR STATION
PENSACOLA, FLORIDA

JOINT PROJECT NM 18 02 99, SUBTASK 1
Report No. 68

The Ohio State University Research Foundation
Columbus, Ohio, under Contract N6ONR 22525
Office of Naval Research Project Designation No. NR 145-993

and

U. S. Naval School of Aviation Medicine

A RATING SCALE TECHNIQUE FOR THE MEASUREMENT OF SPEAKER INTELLIGIBILITY

Report by

Robert W. Peters

Approved by

John W. Black
Project Director

and

Captain Ashton Graybiel, MC, USN
Director of Research

Released by

Captain Julius C. Early, MC, USN
Commanding Officer

8 February 1957

TABLE OF CONTENTS

SUMMARY PAGE

THE PROBLEM

An evaluation was made of the method of measuring speaker intelligibility by listener ratings of voice samples on an equal-appearing intervals scale. Twenty-four speakers and seven panels of listeners, with a minimum of 20 persons in each panel, were involved in the experiment. Recordings were made of each speaker reading multiple-choice intelligibility test word lists and prose material. Ten-second voice samples were prepared from the prose reading. The multiple-choice test material was played for listener panels to provide for each speaker a percent intelligibility score. The ten-second voice samples were played for listening panels under various listening conditions to provide for each speaker a scale value intelligibility score. These listening conditions were that of hearing the voice signal in quiet and at the S/N ratios of +5 db, 0 db, and -5 db. Correlation coefficients were determined between multiple-choice and scale value scores to provide an estimate of the validity of the rating method. An analysis of variance was used to test the significance of the differences among the mean scale values with respect to the different listening conditions.

FINDINGS

Moderately high positive correlations between multiple-choice and scale value intelligibility scores suggest that the rating scale method provides a fairly good estimate of speaker intelligibility. Q-values, which provide an index of reliability, were within reasonable limits. There was a progressive increase in mean scale values as the listening condition became less adverse in the range from -5 db S/N ratio to listening in quiet.

INTRODUCTION

Traditionally, monosyllables, words, and sentences have been used for measuring speaker intelligibility, listener reception, and the efficiency of communication equipment. This has involved speakers reading standardized material and listeners responding to the reading on standardized test forms. The advantages of this type of procedure are many, and it has been through the development and refinement of standardized tests that it has been possible to study voice communication problems extensively.

However, the precision and efficiency of standardized tests has introduced errors and limitations in the measurement of voice communication. A notable departure from the actual communication situation, with resulting errors, is that the speaker is required to read material and, further, to read material which might be quite different from his usual communication transmissions. A major limitation of standardized intelligibility measurement is that systems can be evaluated and experiments conducted only where the speaker can interrupt his activities to read material.

Two examples of problems which cannot adequately be investigated by the standardized tests are 1) the evaluation of actual communication networks, and 2) the effect of stress upon man's communication efficiency. To have operators read standardized material probably gives neither an adequate picture of their efficiency nor the efficiency of the network in which they operate. If a subject in an experiment involving stress were to interrupt his activities to read a series of words, the illusion of stress could hardly be maintained.

To measure intelligibility in the two types of situations suggested above it would be desirable to evaluate actual transmissions. One procedure might be to have listeners write the transmissions and arrive at a ratio score of the number of words correctly reported to the number of words transmitted. A difficulty lies in determining the number of words transmitted. A variation might involve the use of a two-way network and the tabulation of the number of messages that had to be repeated. An alternative method to the above would be to take voice samples from the speaker's transmissions and attempt to assign a quantitative intelligibility value to the samples. This would involve a scaling procedure.

Workers at the Harvard Psycho-Acoustic Laboratory during World War II evaluated the relationship between subjective ratings by judges and intelligibility scores (4). Word and sentence tests were used to provide both the scale and standard intelligibility

measures. The results indicated that valid ratings of intelligibility of talkers can probably be obtained from a small number of trained judges.

In the area of speech pathology, Lewis and Sherman (2) have demonstrated that severity of stuttering can be quantified through the use of a rating technique based on nine-second samples of speech.

The possibility arises that voice intelligibility may be quantified through the use of rating scales to a sufficient extent to be used as a measuring device in problems and experiments where standardized intelligibility tests are not applicable.

The purpose of the present experiment was to evaluate the technique of measuring speaker intelligibility through listener ratings of voice samples on an equal-appearing intervals scale for validity, reliability, and the effect of various S/N ratios upon mean scale values.

METHOD

SUBJECTS

The subjects were drawn from a population of students in the naval aviation flight training program.

TEST MATERIALS

The test materials used to measure speaker intelligibility were Forms A, B (1), A-1, and B-1 (3) of the multiple-choice intelligibility tests and ten-second samples of speakers reading prose material. The prose material read by the speakers was taken from current magazine articles. Twenty-four speakers, also drawn from a population of students in the flight training program, read for the recording of these materials. Each speaker read two word lists from the multiple-choice intelligibility tests and three minutes of prose material. The particular multiple-choice word lists read by each speaker were randomly determined with the restriction that one list for each speaker be either Forms A or B and the other list be from either Forms A-1 or B-1. Four ten-second samples were prepared from the prose read by each speaker. These samples were programmed into a continuous tape with an identifying carrier number preceding each sample. The order of the samples was randomized with the restrictions that each speaker be heard once in each sequential group of 24 samples and that the same voice not appear in adjacent samples.

APPARATUS

The readings by the speakers were recorded on an Ampex, Model 400, magnetic tape recorder fed by an Altec-Lansing, Model 21-C, condenser microphone. Visual monitoring of a VU meter was done to insure relatively the same level for all speakers. The playback equipment for the presentation of this material to the listeners included the Ampex recorder, an Altec-Lansing Model 250-A, control console with an associated line amplifier which fed a headset listening circuit of PDR-3 (Permaflux) receivers. The design of the experiment required that noise be mixed with the voice signal at several S/N ratios. The noise was produced by an H. H. Scott, Model 810-A, noise generator with the control set to produce ASA type white noise.

PROCEDURE

The subjects participated as members of listening panels. There were seven panels of listeners with a minimum of 20 persons in each panel. The task for members of two of the panels was to respond to multiple-choice intelligibility test words. One of these panels responded to the words of Forms A and B and the other panel to the words of Forms A-1 and B-1. These listeners heard the voice signal at approximately 95 db (re 0.0002 dyne/cm^2) with white noise mixed with the signal at a 0 db S/N ratio.

The listeners of the other five panels rated the voice samples of the speakers' readings of prose material on a seven point scale. Four of the panels rated voice samples for intelligibility. Listeners of the fifth panel rated the voice samples in terms of the certainty with which they had understood what was said in the ten-second sample.

With respect to judgments of intelligibility, the scale extended from one, representing least intelligibility, to seven, representing most intelligibility. The listeners heard recorded instructions about the procedures for judging. (See Appendix A.) Included in the instructions were three sets of voice samples arranged in seven steps from least to most intelligible. These three sets of voice samples were judged by four pre-experimental observers to represent seven steps from least to most intelligible and were to assist the listeners in establishing a range of intelligibility. These demonstration samples were prepared by selective low-pass filtering of voice samples read by a single speaker. This speaker was not one of the 24 used in the experiment. The listeners rated 30 voice samples for practice before rating the test samples. These 30 samples were taken from the prose material recorded by the 24 speakers of the experiment.

The listeners who made intelligibility ratings heard the voice signal through their earphones at approximately 95 db. The listening conditions for the four panels differed in that one panel heard the signal in quiet, another with noise mixed with the signal at a +5 db S/N ratio, another at the S/N ratio of 0 db, and a final panel heard the signal at a -5 db S/N ratio. The S/N ratios were achieved by altering the noise level relative to a constant voice signal level.

The panel of listeners who rated the voice samples for certainty of understanding heard the voice signal at approximately 95 db at a 0 db S/N ratio. These listeners also heard recorded instructions indicating how they were to make their judgments (Appendix B) and rated 30 practice samples.

Median scale values and Q-values were determined for the voice samples according to the manner described by Thurstone and Chave (5). For each of the 24 speakers there were both scale value and percent value estimates of intelligibility. The former was provided by scale ratings and the latter by the multiple-choice tests. Each panel of listeners rated each of the 24 speakers four times. The basic scale intelligibility score for each speakers was the mean of these four scale values. Each speaker's multiple-choice intelligibility score was based on listener responses to the two lists read by each speaker.

The experiment was concerned with two aspects of intelligibility scaling: One concerned an estimate of the validity of the method; the other concerned the effects of S/N ratio upon mean intelligibility scale values. Correlation coefficients were determined between scale and multiple-choice values to provide estimates of validity. Scale-value data were treated with analysis of variance to evaluate the effect of S/N ratio upon listener ratings of voice samples.

RESULTS

CORRELATIONS BETWEEN MULTIPLE-CHOICE AND SCALE VALUE INTELLIGIBILITY MEASURES

Product-moment correlations were determined between the speaker multiple-choice intelligibility values and each of the five sets of speaker intelligibility scale values. Since the multiple-choice words were heard by the listeners at a 0 db S/N ratio, the correlations between multiple-choice and the two other sets of speaker scores earned under a 0 db S/N ratio were of primary interest. These were the ratings of

intelligibility and certainty at a 0 db S/N ratio. The correlation between these intelligibility values and multiple-choice values was +.58; that between certainty and multiple-choice was +.58; certainty and intelligibility scale values correlated +.99. Similar correlations as those reported above were computed between multiple-choice scores and the +5 db, -5db, and the quiet intelligibility rating values. These were +.67, +.57, and +.45, respectively.

The scale values for each speaker, based on the first rating of the four ratings made by the listeners of each speaker's voice samples, were correlated with multiple-choice values to provide an estimate of the validity of a single and initial intelligibility rating. These correlations are comparable to the ones reported in the preceding paragraph. The correlations with multiple-choice values were as follows: certainty ratings, +.51; 0 db S/N ratio, +.49; +5 db S/N, +.70; -5 db S/N ratio, +.55; and ratings in quiet, +.48.

The correlations between the multiple-choice and intelligibility scale values probably were attenuated because of errors of measurement in both tests. An estimate of correlation was made with correction for attenuation between the multiple-choice values and the 0 db S/N ratio intelligibility rating values.* Correlations were determined between multiple-choice Forms A and B, and A-1 and B-1 and bewteen first and second, and third and fourth ratings made by the listeners of the voice samples. These correlations were +.78 and +.58, respectively. The estimated correlation between the two tests, corrected for attenuation, was +.84.

An estimate of reliability of the scaling technique is provided by the +.58 correlation between first and second, and third and fourth ratings reported above and by the mean Q-values. The mean Q-values were 0.99 for -5 db S/N ratio, 1.06 for 0 db S/N ratio, 1.06 for +5 db S/N ratio, 1.81 for in quiet rating, and 1.29 for certainty of understanding rating.

THE EFFECT OF S/N RATIO UPON INTELLIGIBILITY SCALE VALUES

Speaker scale values with respect to intelligibility ratings in quiet and at the S/N ratios of +5 db, 0 db, and -5 db were treated with analysis of variance to evaluate the effect of S/N ratio upon mean scale values. Results of the F-test, as summarized in Table I, indicate significant differences among the various listening conditions.

---

* r (corrected for attenuation) = $\frac{r_{xy}}{\sqrt{r_{xx}\, r_{yy}}}$

TABLE I

Summary of an Analysis of Variance Testing Differences Among Four Listening Conditions with Respect to Mean Speaker Intelligibility Scale Values

Source of Variation	df	ss	ms	F	$F_{.05}$
Conditions (C)	3	68.88	22.96	18.52*	2.71
Within-Groups (w)	92	114.47	1.24		
Total	95	183.35			

*$F = ms_C / ms_w$

The mean speaker intelligibility scale values were 3.76, 3.05, 2.48, and 1.45 for the in quiet, +5 db S/N ratio, 0 db S/N ratio, and -5 db S/N ratio listening conditions, respectively. The difference required between means for significance at the five percent level was .63.* It may be noted that mean intelligibility scale values progressively decreased as the listening conditions became increasingly adverse. The only difference between means which was not significant in this progression was the difference between the means for the +5 db and the 0 db S/N ratio listening conditions.

DISCUSSION

The results would seem to indicate that moderately valid estimates of speaker intelligibility may be obtained by scaling by the technique of equal-appearing intervals. The correlations between speaker multiple-choice and the several scale intelligibility values ranged between +.45 and +.67. The lowest correlation was between ratings in quiet and multiple-choice values. The ratings in quiet were somewhat unstable as reflected by the high Q-values, 1.81. However, considering that a homogeneous group of speakers was used in this experiment, this is not particularly surprising. Under the favorable condition of listening in quiet it is understandable that the listeners had difficulty in assigning intelligibility ratings to the voice samples.

To the extent that Q-values are indicative of reliability, the mean Q-values for the other listening conditions are within acceptable limits. The possible exception was the Q-value of 1.29 for the ratings of certainty of understanding.

*Critical difference (d.d.) $= t_{.05} (2ms_w / n)^{1/2} = .63$

Mean scale values reflected the listening conditions under which the voice samples were heard. This was indicated by the progressive increase in mean values as the listening conditions became more favorable in the range from -5 db S/N ratio to listening in quiet. The influence of different S/N ratios upon scale values is encouraging. It suggest that this technique of measuring intelligibility has wider applications than that of evaluating individual differences among a group of speakers.

The correlation of +.99 between certainty of understanding and intelligibility scale values indicates that the two methods are measuring the same factors. Certainty of understanding would be a more desirable criterion for measuring communication efficiency than would intelligibility rating because it would eliminate the need to train listeners to make judgments in keeping with pre-determined levels of intelligibility. The questionable aspect of the certainty judgments was that the Q-value for this measure was somewhat higher than were the Q-values for intelligibility ratings.

An over-all evaluation of the rating scale technique for determining voice intelligibility, as used in this experiment, would suggest that the method has possibilities for measuring intelligibility in problems where the use of standardized intelligibility measures is not feasable. Further evaluation should probably be made of the technique of instructing listeners to make judgments of certainty of understanding. If this technique does not appear promising, then it would be necessary to develop a scale of intelligibility to use in the instruction of listeners who are to make intelligibility judgments.

To estimate the validity of a proposed test by correlating it with established tests is open to legitimate question. Although this was done in this experiment, the purpose was to provide a preliminary estimate of the validity. The measures of intelligibility obtained by a rating scale technique should be validated against other measures of communication efficiency. Perhaps a study comparing scaled estimates of intelligibility of voice samples with write-down intelligibility measures of the same samples would provide a good indication of the validity of the scaling technique.

CONCLUSIONS

The purpose of the present experiment was to evaluate the technique of measuring speaker intelligibility through listener ratings of voice samples on an equal-appearing intervals scale. The technique was evaluated for validity, reliability, and the effect of different S/N ratios upon mean scale values.

The results indicate that the scaling technique provides a fairly good estimate of speaker intelligibility. The rating scale values of intelligibility appear to be reasonably reliable and are influenced by the listening conditions under which the ratings are made by listeners. The method appears to have promise for measuring intelligibility in situations where standardized intelligibility measures are not applicable.

REFERENCES

1. Haagen, C. H., Intelligibility measurement: twenty-four word multiple-choice tests. OSRD Rept. No. 5567, New York, N. Y.: Psychological Corp., 1945.

2. Lewis, D. and Sherman, D., Measuring the severity of stuttering. J. Speech Hearing Dis. 16: 320-326, 1951.

3. Scalero, A. M., Alternate speaker lists for multiple-choice intelligibility tests. Unpublished Master's Thesis, Ohio State Univ., 1953.

4. Stevens, S. S., Abrams, M. H., Goffard, S. J., and Miller, J., Subjective ratings of intelligibility of talkers in noise in "Speech in noise: a study of the factors determining its intelligibility." OSRD Rept. No. 4023. Harvard Psycho-Acoustic Laboratory, 1944.

5. Thurstone, L. L. and Chave, E. J., The Measurement of Attitude. Chicago: University of Chicago Press, 1929.

APPENDIX A

INSTRUCTIONS TO LISTENERS CONCERNING A SEVEN-POINT RATING SCALE OF VOICE INTELLIGIBILITY SAMPLES

APPENDIX A

You are going to rate a series of speech samples for voice intelligibility. Intelligibility relates to how well you understand the voice signal. You are to judge each voice sample in relation to a seven point scale.

The scale is one of equal steps with 1 representing the least intelligible signal and 7 representing the most intelligible signal. Step 4 is halfway between 1 and 7. Do not attempt to make any of your judgments between any two of these seven points but only at these points.

Each voice saying the samples is repeated several times. You may thus recognize that you have previously rated a certain voice. However, make an attempt to give an independent rating to the voice sample each time this occurs.

Now you will hear a series of voice samples which will help you establish range of voice intelligibility for the purpose of making your ratings. The samples are arranged in order of least to most intelligible.

Here is another series of voice samples ranging from least to most intelligible.

The following is still another series of voice samples ranging from least to most intelligible.

You will now hear the series of voice samples to be judged. Remember to judge each of the samples on the seven point scale with 1 representing the least intelligible and 7 representing the most intelligible. Step 4 is thus halfway between 1 and 7 in intelligibility with the other points falling on the scale equal distances apart. Do not attempt to place the samples between any two of the seven points, but only at these points.

The first thirty samples are to be judged for practice and to further acquaint you with the range of intelligibility among these samples.

APPENDIX B

INSTRUCTIONS TO LISTENERS CONCERNING A SEVEN-POINT RATING SCALE OF CERTAINTY OF UNDERSTANDING

APPENDIX B

You are going to rate a series of voice samples according to how certain you are that you have understood the sample. You are to judge each voice sample in relation to a seven point scale.

The scale is one of equal steps with 1 representing the least certainty and 7 representing the most certainty that you have understood the voice sample. Step 4 is halfway between 1 and 7. Do not attempt to make any of your judgments between any two of these seven points but only at these points.

Each voice saying the samples is repeated several times. You may thus recognize that you have previously rated a certain voice. However, make an attempt to give an independent rating to the voice sample each time that this occurs.

Remember to rate each of the samples according to how certain you are that you understood the sample on the seven point scale with 1 representing least certainty and 7 representing most certainty. Step 4 is thus halfway between 1 and 7 in certainty with the other points falling on the scale equal distances apart. Do not attempt to place the samples between any two of the seven points but only at these points.

They first thirty samples are to be judged for practice and to further acquaint you with the range of certainty of your judgments among these samples.